Subtree power analysis finds optimal species for comparative genomics

Jon D. McAuliffe,[1] Michael I. Jordan,[1,2] Lior Pachter[3,*]

[1]Department of Statistics, [2]Computer Science Division, [3]Department of Mathematics,
University of California,
Berkeley, CA 94720, USA

[*]To whom correspondence should be addressed; E-mail: lpachter@math.berkeley.edu.

Abstract

Sequence comparison across multiple organisms aids in the detection of regions under selection. However, resource limitations require a prioritization of genomes to be sequenced. This prioritization should be grounded in two considerations: the lineal scope encompassing the biological phenomena of interest, and the optimal species within that scope for detecting functional elements. We introduce a statistical framework for optimal species subset selection, based on maximizing power to detect conserved sites. In a study of vertebrate species, we show that the optimal species subset is not in general the most evolutionarily diverged subset. Our results suggest that marsupials are prime sequencing candidates.

Introduction

Comparative genomic methods can reveal conserved regions in multiple organisms, including functional elements undetected by single-sequence analyses [1, 2]. Individual studies have demonstrated the effectiveness of genomic comparison for specific regions and elements [3, 4, 5, 6, 7]. Such successes indicate that comparative considerations should play a major role in decisions about what unsequenced species to sequence next. For comparative purposes, sequencing choices must first of all be guided by specification of the widest range of species sharing the functions or characters in question, which we call the lineal scope [8] [1]. Boffelli *et*

[1]This differs from the "phylogenetic scope" of Cooper *et al.* [9] in that lineal scope is determined solely by a biological trait of interest, whereas phylogenetic scope can be determined according to any considerations.

al. [10] discuss the utility of comparisons in lineal scopes ranging from the primate clade to the vertebrate tree.

Most lineal scopes selected in practice will include far more extant species than can be sequenced with today's resources. Thus, sequencing prioritization is an unavoidable issue, both for smaller-scale efforts targeting particular regions and for whole-genome projects, whose focus should reflect in part the aggregate needs of comparative analyses. Few studies on comparative methods provide a quantitative framework for decision-making about what to sequence. An exception is the work of Sidow and others [9, 11]: given a set of sequenced organisms and an inferred phylogeny, Cooper *et al.* [9] argue that decisions should be based on maximizing additive evolutionary divergence in a phylogenetic tree.

While additive divergence captures part of the problem underlying organism choice, it fails to reflect the inherent tradeoff that characterizes the problem. On the one hand, the success of procedures for assessing conservation does depend on sufficient evolutionary distance among the sequences [5, 4, 12]. On the other hand, a given set of species may have diverged too far from one another to be useful, even when orthology is preserved: in the limit of large evolutionary distance, conservation and nonconservation are just as indistinguishable as at distance zero [13]. Furthermore, phylogenetic topology has counterintuitive effects on usefulness.

Here, we present a decision-theoretic framework which subsumes these issues, providing a procedure for making systematic, quantitative choices of species to sequence. Statistical power is our optimality criterion for species selection. Thus, we measure the effectiveness of a species subset directly in terms of error rates for detecting and overlooking conservation at a single orthologous site. Measuring power disentangles effects due to the number of species used from effects due to relative evolutionary distances in the phylogeny. We illustrate these ideas theoretically, in a star phylogeny analysis, and practically, with an empirically-derived phylogeny on 21 representative vertebrate species. The results indicate that adding the dunnart or a closely-related marsupial to finished and underway vertebrate sequences would most increase the power to detect conservation at single-nucleotide resolution.

Setup

We frame conservation detection in the following decision-theoretic setting. The data $\mathbf{x}$ are the nucleotides at an orthologous site across a set of species, i.e., an ungapped alignment column. We view these bases as corresponding to the leaves of a phylogeny with unobserved ancestral bases. We assume that the phylogenetic topology, the Markov substitution process along the branches, and the branch lengths are all known. The phylogeny induces the observed-data probability distribution $p(\mathbf{x}; r)$ as the marginal distribution on its leaves, which can be evaluated efficiently for any $\mathbf{x}$ and r [14]. The parameter $r > 0$ is an unknown global mutation rate shared among all branches. We choose two threshold values $r_N > r_C$ for r: an actual mutation rate of at least r_N corresponds by definition to a nonconserved site, whereas a rate no more than r_C means the site is strongly conserved. When $r_N > r > r_C$, the conservation is too weak to

interest us.

The decision-theoretic goals are now twofold. First, fixing a set of species, we wish to select a decision rule $\delta(\mathbf{x})$ which declares the site either nonconserved ($\delta(\mathbf{x}) = 0$) or conserved ($\delta(\mathbf{x}) = 1$) using only data from those species. Every nontrivial $\delta(\mathbf{x})$ will have positive probability of making two mistakes: when $r \geq r_N$, $P_r(\delta(\mathbf{X}) = 1)$ is the probability it erroneously detects conservation, and when $r \leq r_C$, $P_r(\delta(\mathbf{X}) = 0)$ is the probability it overlooks conservation. Minimizing these probabilities guides our choice of $\delta(\mathbf{x})$. We formulate a Neyman-Pearson hypothesis test [15] of the null hypothesis H_0: $r \geq r_N$ versus the alternative hypothesis H_A: $r \leq r_C$, stipulating a maximum allowed probability α of falsely rejecting H_0 (falsely declaring conservation). Controlling this error probability is a central concern [9]. Subject to this constraint, we find a test statistic $\delta(\mathbf{x})$ with large power to detect conservation, that is, small probability of overlooking conservation. The second goal is to maximize this power over subtrees in the larger phylogeny determined by the chosen lineal scope, such as all subtrees on k extant species within the anthropoid clade, where k is determined by sequencing resource limitations.

Symmetric star topology

We initially pursue these goals in a phylogenetic setting called the symmetric star topology (SST), where k extant species are connected to a single ancestor by branches of common length $t > 0$. Choosing k and t in the SST is akin to choosing k extant species within a larger phylogeny, such that each pair of chosen species is at a distance of approximately $2t$. Hypothesis testing in the fully-observed SST (FOSST), with known ancestral base, closely approximates testing in the hidden-ancestor SST (HASST), the case of interest, for small to moderate t (Figure 1). This follows because there is little uncertainty about the ancestral base at short evolutionary distances: with high probability, it equals the most-occurring base among the descendants. The analogy matters because we know the uniformly most-powerful testing procedure under the FOSST: it rejects H_0 (declares conservation) for large values of the likelihood ratio statistic $p(\mathbf{x}; r_C)/p(\mathbf{x}; r_N)$ (see Appendix).

Figure 1A shows the power of the FOSST likelihood-ratio test against the particular alternative distribution $r = r_C$, as t and k vary. Power against other alternatives $r < r_C$ is larger (see Appendix). For each t, power increases monotonically in k. However, for each k, there is a unique power-maximizing branch length $t^*(k)$. In the Appendix we explain this in terms of stationary Markov substitution processes. Fundamentally, it happens because both nonconserved and conserved sites accrue mutations, and the difference in their mutation rates becomes irrelevant as $t \to \infty$. A consequence of this is the suboptimality of maximizing additive divergence: for any k, the optimal tree has finite divergence $k \cdot t^*(k)$, rather than arbitrarily large divergence. Comparing Figures 1A and 1B shows the FOSST accurately approximates the HASST in a large interval around $t^*(k)$ for $k > 2$, so the conclusion also applies to the HASST. As k increases, $t^*(k)$ stabilizes at a nonzero value (Figure 2). Thus, the optimal divergence $k \cdot t^*(k)$ grows

without bound as a function of k.

Empirical power analysis

We now explore subtree power maximization empirically, using the previously-reported CFTR sequence data [6] on 21 representative vertebrates (Table 1). We estimated a phylogeny (Figure 3) based on a multiple sequence alignment, as described in the Appendix. This procedure yields phylogenies applicable to data outside the estimation region [9, 16]. We formulated the likelihood-ratio statistic and calibrated the conserved rate threshold r_C to correspond to typical genic conservation in the sequenced region. Having fixed the form of the testing procedure, the goal is to maximize its power to detect conservation over subsets of size k chosen from among the 21 species, for various values of k. This entails searching for the maximal-power family subtree, or k-most-powerful Steiner subtree (k-MPSS), among the $\binom{21}{k}$ subtrees with k leaves (see Appendix). A Steiner subtree on k leaves is the unique smallest subtree rooted at their last common ancestor.

Table 2 shows the k-MPSS (starred) in comparison to the subtree on k leaves with largest additive divergence (the k-most-divergent Steiner subtree, or k-MDSS, daggered). The latter has been the focus of previous work [4, 17, 9]. These two subtree selection criteria do not coincide. For instance, at $r_N = 2$, the 5-MPSS includes the dunnart, whereas the 5-MDSS instead uses the platypus. The t-statistic on the difference in power is 2.06, so variability in the power estimate is not a likely explanation. A more extreme example is $r_N = 10$: the 4-MPSS and 4-MDSS have only one species in common, and the absolute loss in power that results from using the 4-MDSS is nearly 8.5% (t-statistic 105.7). Here, more than 4,400 subtrees have higher power than the 4-MDSS. The disagreement at higher values of k underscores the effect of phylogenetic topology on the detection of conservation.

We carried out a similar comparison, under the constraint that the 9 completely or partially sequenced vertebrates in the data set appear in the subtree (Table 3). This reveals the species whose addition to the current sequencing mix would most improve the power to detect single-site conservation. As in Table 2, the most-powerful and most-divergent subtrees generally differ. The pattern of disagreement is not systematic: when $r_N = 5$, for example, they disagree at 10 and 11 species, agree at 12 and 13, and disagree at 14. Table 2 exhibits similar properties. We conclude that the k-MDSS is not a reliable surrogate for the k-MPSS. Table 3 reveals that the single most beneficial species to sequence next is the dunnart (improving power by a relative 12.5%), whereas the species which adds the most evolutionary divergence is the platypus.

Discussion

Even when the MPSS and MDSS coincide, the decision-theoretic point of view puts the focus on the important issue: the two kinds of discrimination errors and their probabilities. The power

calculation directly measures the marginal benefit of additional sequenced species as an increase in probability of conserved site detection. This enables us to choose a k which optimizes the tradeoff between the expected benefit of detecting conservation and the cost of additional sequencing. By contrast, the additive divergence of a species set gives no direct indication of how a procedure using those species will fare. Since the phylogeny and substitution process are parameters of our procedure, their choice can be tailored to particular investigations. Our emphasis on single-site detection of conservation will lead to conservative power estimates in situations where conservation is tested for simultaneously across multiple correlated sites.

Acknowledgments

We thank Peter Bickel and Adam Siepel for helpful comments. L.P. was supported by a grant from the NIH (R01-HG2362-3), a Sloan Foundation Research Fellowship, and an NSF Career award (CCF-0347992).

Appendix

Symmetric star topology

Fully observed

Let $\mathcal{P} = \{p(x_0, \mathbf{x}; r) : r > 0\}$ be the family of FOSST probability mass functions indexed by the mutation rate parameter r, for some fixed choice of descendant species count k and common branch length t. Here x_0 is the observed ancestral base and $\mathbf{x} = (x_1, \ldots, x_k)$ are the observed descendant bases. Write

$$n(x_0, \mathbf{x}) = \sum_{i=1}^{k} \delta(x_0, x_i) \; , \tag{1}$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta function. Under the Jukes-Cantor substitution process, with its equilibrium distribution (the uniform distribution) on the ancestral base, each member of $\mathcal{P}$ has the form

$$\begin{aligned} p(x_0, \mathbf{x}; r) &= \frac{1}{4} \prod_{i=1}^{k} \left(\frac{1 + 3e^{-4rt}}{4} \right)^{\delta(x_0, x_i)} \left(\frac{3(1 - e^{-4rt})}{4} \right)^{1 - \delta(x_0, x_i)} & (2) \\ &= \frac{1}{4} \left(\frac{1 + 3e^{-4rt}}{4} \right)^{\sum_{i=1}^{k} \delta(x_0, x_i)} \left(\frac{3(1 - e^{-4rt})}{4} \right)^{k - \sum_{i=1}^{k} \delta(x_0, x_i)} \; . & (3) \end{aligned}$$

Fixing $r_C = 1$ entails no loss in generality, due to the nonidentifiability of the parameter pair (r, t) in the Jukes-Cantor substitution process. Choose $r_N > 1$. Substituting (1) into (3) and

simplifying the ratio $p(x_0, \mathbf{x}; 1)/p(x_0, \mathbf{x}; r_N)$ shows that the likelihood-ratio statistic for testing $H_0 : r \geq r_N$ versus $H_A : r \leq 1$ in the FOSST model has the form

$$\frac{(1+3e^{-4t})^{n(x_0,\mathbf{x})}(1-e^{-4t})^{k-n(x_0,\mathbf{x})}}{(1+3e^{-4r_N t})^{n(x_0,\mathbf{x})}(1-e^{-4r_N t})^{k-n(x_0,\mathbf{x})}} \; . \tag{4}$$

The family $\mathcal{P}$ has a monotone (decreasing) likelihood ratio in the statistic $n(x_0, \mathbf{x})$, because for each pair of rate parameters $r_1 > r_2 > 0$, the likelihood ratio

$$\frac{(1+3e^{-4r_1 t})^{n}(1-e^{-4r_1 t})^{k-n}}{(1+3e^{-4r_2 t})^{n}(1-e^{-4r_2 t})^{k-n}} = \left(\frac{1+3e^{-4r_1 t}}{1+3e^{-4r_2 t}}\right)^{n} \left(\frac{1-e^{-4r_1 t}}{1-e^{-4r_2 t}}\right)^{k-n} \tag{5}$$

is a decreasing function of $n = n(x_0, \mathbf{x}) \in \{0, 1, \ldots, k\}$. This follows upon observing that, when $r_1 > r_2$,

$$\frac{1+3e^{-4r_1 t}}{1+3e^{-4r_2 t}} < 1 \quad \text{and} \quad \frac{1-e^{-4r_1 t}}{1-e^{-4r_2 t}} > 1 \; .$$

Standard monotone likelihood-ratio theory [15] therefore implies that the likelihood-ratio test $\mathcal{T}_\alpha$, which rejects when (4) exceeds a critical value c_α, is uniformly most powerful for testing $H_0 : r \geq r_N$ versus $H_A : r \leq 1$ at size α. The size is attained at the particular null distribution $r = r_N$.

The theory also implies that, among the alternative distributions in H_A, $\mathcal{T}_\alpha$ attains its lowest power against $r = 1$, yielding a lower bound on the power against any member of H_A. The power of $\mathcal{T}_\alpha$ against the particular alternative $r = 1$ can be written explicitly as a function of k and t:

$$\rho(k, t) = G_A(n_\alpha + 1; k) + \left(\frac{\alpha - G_0(n_\alpha + 1; k)}{f_0(n_\alpha; k)}\right) f_A(n_\alpha; k) \; . \tag{6}$$

Here, $f_0(\cdot; k)$ and $f_A(\cdot; k)$ are the probability mass functions of a $\mathrm{Bin}(k, d(r, t))$ random variable with $r = r_N$ and $r = 1$, respectively; $G_0(\cdot; k)$ and $G_A(\cdot; k)$ are the corresponding (cadlag) cumulative binomial right-tail probabilities; $d(r, t) = (1 + 3\exp(-4rt))/4$; and n_α is a known critical value. To derive (6), first note that $\mathcal{T}_\alpha$ is equivalent to the test which rejects H_0 when the statistic $n(x_0, \mathbf{x})$ exceeds a corresponding critical value n_α, again by virtue of the monotone likelihood-ratio property. Both tests thus have the same power $\rho(k, t)$. Let P_0 and P_A denote the distribution of $n(X_0, \mathbf{X})$ under $r = r_N$ (the size-determining distribution) and $r = 1$, respectively. Because $n(x_0, \mathbf{x})$ can take on only finitely many values, we use randomized rejection to achieve level exactly α. The critical value is $n_\alpha = \min\{n : P_0(n(X_0, \mathbf{X}) > n) \leq \alpha\}$. When $n(x_0, \mathbf{x}) > n_\alpha$, we reject. When $n(x_0, \mathbf{x}) = n_\alpha$, we reject with probability $\gamma(\alpha)$ satisfying

$$P_0(n(X_0, \mathbf{X}) > n_\alpha) + \gamma(\alpha) P_0(n(X_0, \mathbf{X}) = n_\alpha) = \alpha \; . \tag{7}$$

This implies that setting

$$\gamma(\alpha) = \frac{\alpha - P_0(n(X_0, \mathbf{X}) > n_\alpha)}{P_0(n(X_0, \mathbf{X}) = n_\alpha)} \tag{8}$$

guarantees a test with size α. It now follows that

$$\rho(k,t) = P_A(n(X_0,\mathbf{X}) > n_\alpha) + \gamma(\alpha)P_A(n(X_0,\mathbf{X}) = n_\alpha)\ . \tag{9}$$

Under the star topology and Jukes-Cantor substitution process, each descendant nucleotide X_i has probability $d(r,t) = (1 + 3\exp(-4rt))/4$ of differing from X_0, independent of all other descendants. Thus $n(X_0,\mathbf{X})$ is a $\mathrm{Bin}(k, d(r,t))$ random variable. Equation (6) follows upon substituting $G_0(n_\alpha + 1; k)$ for $P_0(n(X_0,\mathbf{X}) > n_\alpha)$, $f_0(n_\alpha; k)$ for $P_0(n(X_0,\mathbf{X}) = n_\alpha)$, and similarly for P_A.

Equation (6) involves only known constants and binomial probabilities. The latter can be evaluated quickly to desired accuracy [18]. This allows us to compute $\rho(k,t)$ for many choices of k and t, leading to the power curves in Figure 1A. The kinks in each power curve correspond to settings of t at which the critical value of the likelihood-ratio test changes. The locations of the kinks are easily determined, and the power curves are highly smooth between kinks. Thus, we can find $t^*(k)$ and $\rho^*(k)$ rapidly using a numerical optimization routine (Figure 1A, Figure 2A).

Hidden ancestor

Under the HASST model and Jukes-Cantor process, the likelihood-ratio statistic has the form

$$\frac{\sum_{x_0}(1+3e^{-4t})^{n(x_0,\mathbf{x})}(1-e^{-4t})^{k-n(x_0,\mathbf{x})}}{\sum_{x_0}(1+3e^{-4r_N t})^{n(x_0,\mathbf{x})}(1-e^{-4r_N t})^{k-n(x_0,\mathbf{x})}}\ . \tag{10}$$

This is more difficult to deal with than (4). It is clear that (10) depends only the occurrence counts of the four different bases, not on the leaf configuration which gives rise to the counts. Indeed, (10) is invariant when the bases associated with the counts are permuted. This means that there are only as many distinct values of (10) as there are integer partitions of k into four parts, with partition values of zero allowed. The number of leaf configurations corresponding to each integer partition is an easy combinatorial quantity. We can generate all the integer partitions and evaluate the HASST probability mass function at each one quickly, even for k as large as 100.

Together, these facts allow us to compute the exact null distribution $r = r_N$ and alternative distribution $r = 1$ of (10), for each required setting of (α, r_N, k, t). This yields the power of the HASST likelihood-ratio test, using formulas (8) and (9) with the HASST distribution functions substituted for P_0 and P_A. We then maximize each curve $\rho(k,\cdot)$ by brute force to determine $t^*(k)$ and $\rho^*(k)$ (Figure 1B, Figure 2B).

Existence of maximal power

We can explain the existence of a power-maximizing common branch length t^* under the Jukes-Cantor process, and more generally under any continuous stationary Markov process, as follows. Fix k. At evolutionary distance zero ($t = 0$), the distribution $p(\mathbf{x}; r)$ in a symmetric star

topology is the same for every mutation rate r. Thus the null and alternative hypotheses coincide. In this circumstance, the power is easily seen to equal α. In the limit of evolutionary time, as $t \to \infty$, the distribution of each descendant base approaches the process's stationary distribution, independent of the ancestral base. Since the stationary distribution does not involve the rate r, all conserved and nonconserved distributions converge to the same limit. The limiting power in t is therefore again α. The fact that power begins at α when $t = 0$ and approaches α as $t \to \infty$, together with the fact that power is continuous in t and greater than α on $(0, \infty)$, implies a maximal power must be attained by some finite $t^*(k)$.

Empirical power analysis

We constructed a multiple alignment of 21 sequences (Table 1) from the CFTR data set [6] using MAVID [19]. We then used maximum likelihood [20, 14] to fit a phylogenetic tree topology and branch lengths (Figure 3) to to the alignment. Both the phylogeny estimation and subsequent power analysis employed the nucleotide substitution process of Felsenstein [21], using a transition-transversion ratio of 2:1 and a uniform equilibrium nucleotide distribution. Branch lengths $\{t_j\}$ are measured in expected number of substitutions at an exonic aligned site.

The phylogenetic topology of Figure 3 differs in a few ways from estimates based on considerations of large-scale indel mutations and morphology, for example in its placement of the chicken and platypus. At issue here, however, is its suitability for a single-site power analysis under a substitutional mutation model. We chose our tree estimation procedure to obtain a phylogeny compatible with the data and directed to this goal.

Finding the k-MPSS in a phylogeny is a combinatorial optimization problem, which we solve directly in small to moderate-sized cases by evaluating the power of the likelihood-ratio test based on every candidate Steiner subtree (Table 2). We can also solve the problem directly for larger k, by constraining the species at many of the leaves in the subtree (Table 3). In order to compute power with a particular subtree, we used a Monte Carlo strategy. For each setting of r_N, with $\alpha = 0.05$, we generated 100,000 realizations from the null ($r = r_N$) and alternative ($r = 1$) distributions on the leaves of the full phylogeny. This induced null and alternative empirical distributions on the leaves of every possible subtree, from which we obtained approximations to the true null and alternative distributions of the likelihood-ratio test. These yielded approximate critical values as well as power estimates. We repeated the process of simulation and subtree power estimation ten times for each parameter setting; Tables 2 and 3 show averages and standard errors across repetitions.

References

[1] R. H. Waterston, R. H. Waterston, R. H. Waterston, R. H. Waterston, R. H. Waterston, and R. H. Waterston. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[2] Rat Genome Sequencing Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428:493–521, 2004.

[3] J. Flint, C. Tufarelli, J. Peden, K. Clark, R. Daniels, R. Hardison, W. Miller, S. Philipsen, K. C. Tan-Un, T. McMorrow, J. Frampton, B. Alter, A. Fischauf, and D. Higgs. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Hum. Mol. Genet.*, 10(4):371–382, 2001.

[4] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299:1391–1394, 2003.

[5] E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S. E. Antonarakis. Evolutionary discrimination of mammalian conserved non-genic sequences. *Science*, 302:1033–1035, 2003.

[6] J. W. Thomas, J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell, B. Maskeri, N. F. Hansen, M. S. Schwartz, R. J. Weber, W. J. Kent, D. Karolchik, T. C. Bruen, R. Bevan, D. J. Cutler, S. Schwartz, L. Elnitski, J. R. Idol, A. B. Prasad, S. Q. Lee-Lin, V. V. Maduro, T. J. Summers, M. E. Portnoy, N. L. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C. P. Brinkley, S. Y. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S. L. Ho, M. C. Huang, E. Karlins, P. L. Laric, R. Legaspi, M. J. Lim, Q. L. Maduro, C. A. Masiello, S. D. Mastrian, J. C. McCloskey, R. Pearson, S. Stantripop, E. E. Tiongson, J. T. Tran, C. Tsurgeon, J. L. Vogt, M. A. Walker, K. D. Wetherby, L. S. Wiggins, A. C. Young, L. H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C. L. Shu, P. J. De Jong, C. E. Lawrence, A. F. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E. D. Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.

[7] M. A. Chapman, I. J. Donaldson, J. Gilbert, D. Grafham, J. Rogers, A. R. Green, and B. Göttgens. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.*, 14:313–318, 2004.

[8] S. J. O'Brien, E. Eizirik, and W. J. Murphy. On choosing mammalian genomes for sequencing. *Science*, 292:2264–2266, 2001.

[9] G. M. Cooper, M. Brudno, NISC Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.*, 13:813–820, 2003.

[10] D. Boffelli, M. Nobrega, and E. M. Rubin. Comparative genomics at the vertebrate extremes. *Nature Rev. Genet.*, 5:456–465, 2004.

[11] A. Sidow. Sequence first. Ask questions later. *Cell*, 111:13–16, 2002.

[12] E. H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13:2507–2518, 2003.

[13] L. G. Zhang, V. Pavlovic, C. R. Cantor, and S. Kasif. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.*, 13:1190–1202, 2003.

[14] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum-likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[15] E. Lehmann. *Testing Statistical Hypotheses*. Springer-Verlag, 2nd edition, 1986.

[16] V. B. Yap and L. Pachter. Identification of evolutionary hotspots in the rodent genomes. *Genome Res.*, 14:574–579, 2004.

[17] J. D. McAuliffe, L. Pachter, and M. I. Jordan. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, 20:1850–1860, 2004.

[18] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1974.

[19] N. Bray and L. Pachter. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.*, 14:693–699, 2004.

[20] G. J. Olsen, H. Matsuda, R. Hagstrom, and R. Overbeek. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, 10:41–48, 1994.

[21] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.

	Species
1	Baboon
2	Cat
3	Chicken
4	Chimpanzee
5	Cow
6	Dog
7	Dunnart
8	Fugu
9	Hedgehog
10	Horse
11	Human
12	Lemur
13	Macaque
14	Mouse
15	Opossum
16	Pig
17	Platypus
18	Rabbit
19	Rat
20	Tetraodon
21	Zebrafish

Table 1: The 21 species whose CFTR region sequence data underlie the empirical power analysis.

r_N	Size	Species: $\star$ = MPSS, † = MDSS	Power% (SE)	t vs. MPSS	Rank
2	2	Rat, Zebrafish $\star$†	6.79 (0.01)		1.3
	3	Rat, Zebrafish, Chicken $\star$†	8.30 (0.01)		1.6
	4	Rat, Zebrafish, Chicken, Dog $\star$†	9.61 (0.02)		3.3
	5	Rat, Zebrafish, Chicken, Dog, Dunnart $\star$	10.88 (0.03)		4.4
		Rat, Zebrafish, Chicken, Dog, Platypus †	10.80 (0.02)	2.06	21.7
5	2	Rat, Zebrafish $\star$†	10.60 (0.02)		3.2
	3	Rat, Zebrafish, Chicken $\star$†	21.61 (0.06)		1.8
	4	Rat, Zebrafish, Chicken, Dog $\star$†	39.33 (0.17)		5.2
	5	Rabbit, Cat, Dunnart, Chicken, Hedgehog $\star$	49.96 (0.07)		12.2
		Rat, Zebrafish, Chicken, Dog, Platypus †	47.31 (0.07)	25.82	3894.4
10	2	Dunnart, Lemur $\star$	13.30 (0.03)		21.0
		Rat, Zebrafish †	12.67 (0.02)	16.67	153.0
	3	Dunnart, Cat, Zebrafish $\star$	37.53 (0.11)		10.4
		Rat, Zebrafish, Chicken †	36.83 (0.12)	4.13	77.2
	4	Dunnart, Chicken, Hedgehog, Opossum $\star$	64.69 (0.05)		4.4
		Rat, Zebrafish, Chicken, Dog †	56.21 (0.06)	105.70	4439.3
	5	Macaque, Lemur, Dog, Cow, Pig $\star$	69.75 (0.11)		8.6
		Rat, Zebrafish, Chicken, Dog, Platypus †	66.86 (0.07)	22.28	4867.4

Table 2: The k-MPSS and k-MDSS as a function of the nonconserved rate r_N and the size k of the subtree, with $\alpha = 0.05$ throughout. Results are across 10 repetitions of the Monte Carlo power estimation procedure (see Appendix). The last three columns display the average power (and standard error), the t-statistic for the power difference between the k-MDSS and the k-MPSS (in cases where they differ), and the average power ranking (among all subtrees). Since r_C is calibrated to exonic conservation, the settings of r_N range from a neutral rate ($r_N = 2$) [16] towards extreme single-site mutability.

r_N	Size	New species: $\star$ = MPSS, † = MDSS	Power% (SE)	t vs. MPSS	Rank
	9	{*clamped species only*}	12.81 (0.03)		
	10	Dunnart $\star$	14.42 (0.04)		1.1
		Platypus †	14.25 (0.04)	2.92	3.4
2	11	Dunnart, Platypus $\star$	16.08 (0.05)		1.6
		Platypus, Hedgehog †	15.85 (0.04)	3.62	6.2
	12	Dunnart, Platypus, Hedgehog $\star$†	17.88 (0.06)		1.5
	13	Dunnart, Platypus, Hedgehog, Rabbit $\star$†	19.80 (0.08)		1.1
	14	Dunnart, Platypus, Hedgehog, Rabbit, Cow $\star$†	21.41 (0.08)		1.6
	9	{*clamped species only*}	56.44 (0.16)		
	10	Dunnart $\star$	65.59 (0.20)		1.0
		Platypus †	64.74 (0.17)	3.18	3.0
	11	Dunnart, Opossum $\star$	71.05 (0.09)		2.3
5		Platypus, Hedgehog †	70.54 (0.06)	4.74	14.6
	12	Dunnart, Platypus, Hedgehog $\star$†	72.77 (0.08)		1.2
	13	Dunnart, Platypus, Hedgehog, Rabbit $\star$†	76.02 (0.13)		1.0
	14	Dunnart, Platypus, Hedgehog, Rabbit, Opossum $\star$	80.41 (0.10)		2.2
		Dunnart, Platypus, Hedgehog, Rabbit, Cow †	80.08 (0.14)	1.88	2.1
	9	{*clamped species only*}	86.61 (0.06)		
	10	Platypus $\star$†	91.67 (0.06)		1.3
	11	Dunnart, Opossum $\star$	94.07 (0.02)		3.3
		Platypus, Hedgehog †	93.96 (0.03)	2.66	10.7
10	12	Dunnart, Platypus, Rabbit $\star$	95.84 (0.03)		2.4
		Dunnart, Platypus, Hedgehog †	95.79 (0.30)	1.30	4.4
	13	Dunnart, Platypus, Rabbit, Opossum $\star$	97.31 (0.02)		4.6
		Dunnart, Platypus, Rabbit, Hedgehog †	97.29 (0.02)	0.85	6.6
	14	Dunnart, Platypus, Rabbit, Hedgehog, Opossum $\star$	97.99 (0.01)		2.4
		Dunnart, Platypus, Rabbit, Hedgehog, Cow †	97.95 (0.02)	1.83	7.6

Table 3: The k-MPSS and k-MDSS, under the constraint that the following nine species are included in the subtree: human, mouse, rat, chimpanzee, dog, chicken, fugu, zebrafish, and tetraodon. The scheme of the table is the same as Table 1.

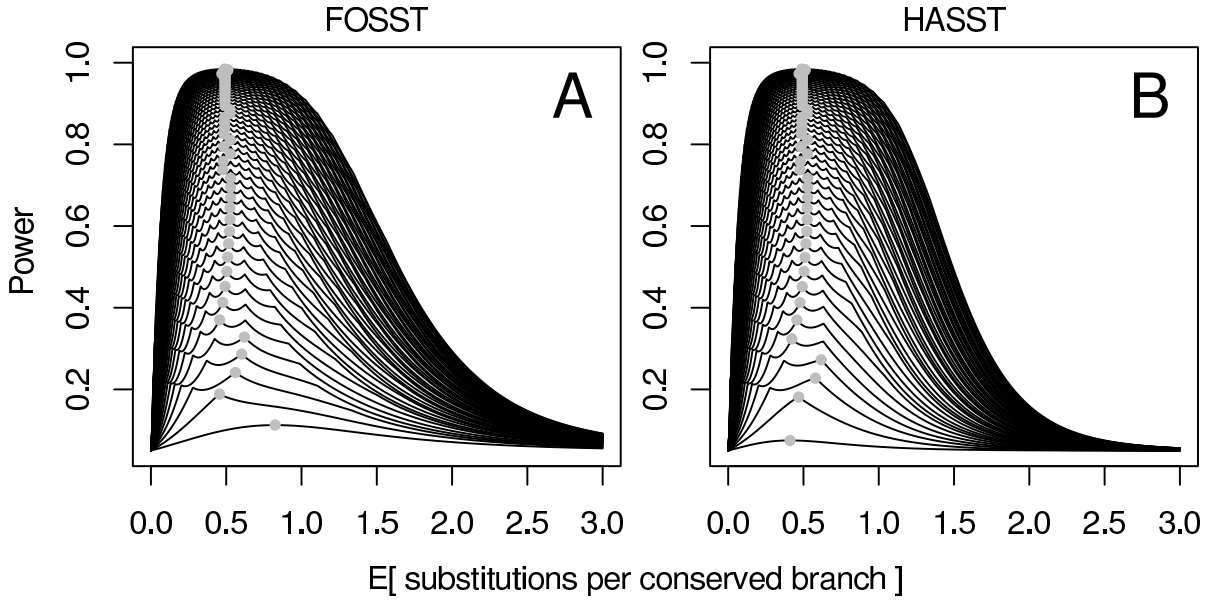

Figure 1: Power to detect conservation as a function of common branch length for the fully-observed (A) and hidden-ancestor (B) SSTs, using $r_C = 1$, $r_N = 2$, and $\alpha = 0.05$. Each power curve corresponds to an even number k of observed descendant species, from two (bottommost curve) to 100 (topmost). The maximum power attained for each k is indicated by a grey dot. The power analysis uses the Jukes-Cantor substitution process; power curves are computed analytically (see Appendix). Power curves computed with other values of r_N and α remain qualitatively the same (not shown).

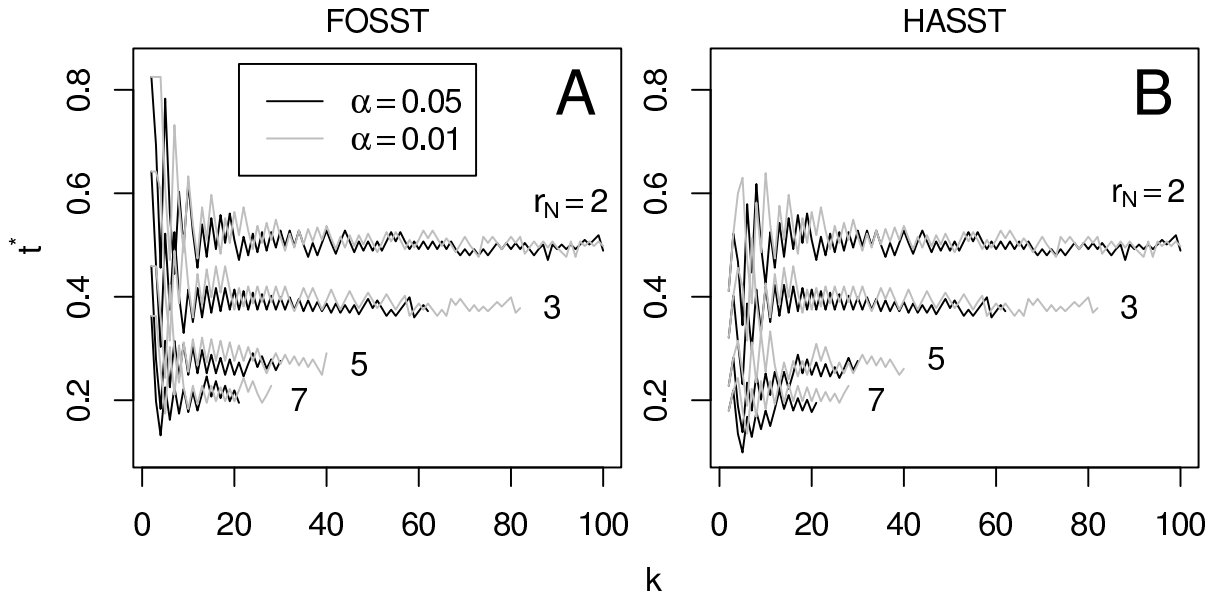

Figure 2: The optimal common branch length $t^*(k)$ in the fully-observed (A) and hidden-ancestor (B) SSTs, as a function of the number of descendant species k. Each black curve uses the indicated nonconserved rate $r_N = 2, 3, 5, 7$ with $\alpha = 0.05$; grey curves are analogous with $\alpha = 0.01$. As k increases, $t^*(k)$ stabilizes at a value depending on r_N but not α. For the larger r_N's, the curves are terminated when power reaches 99.9%.

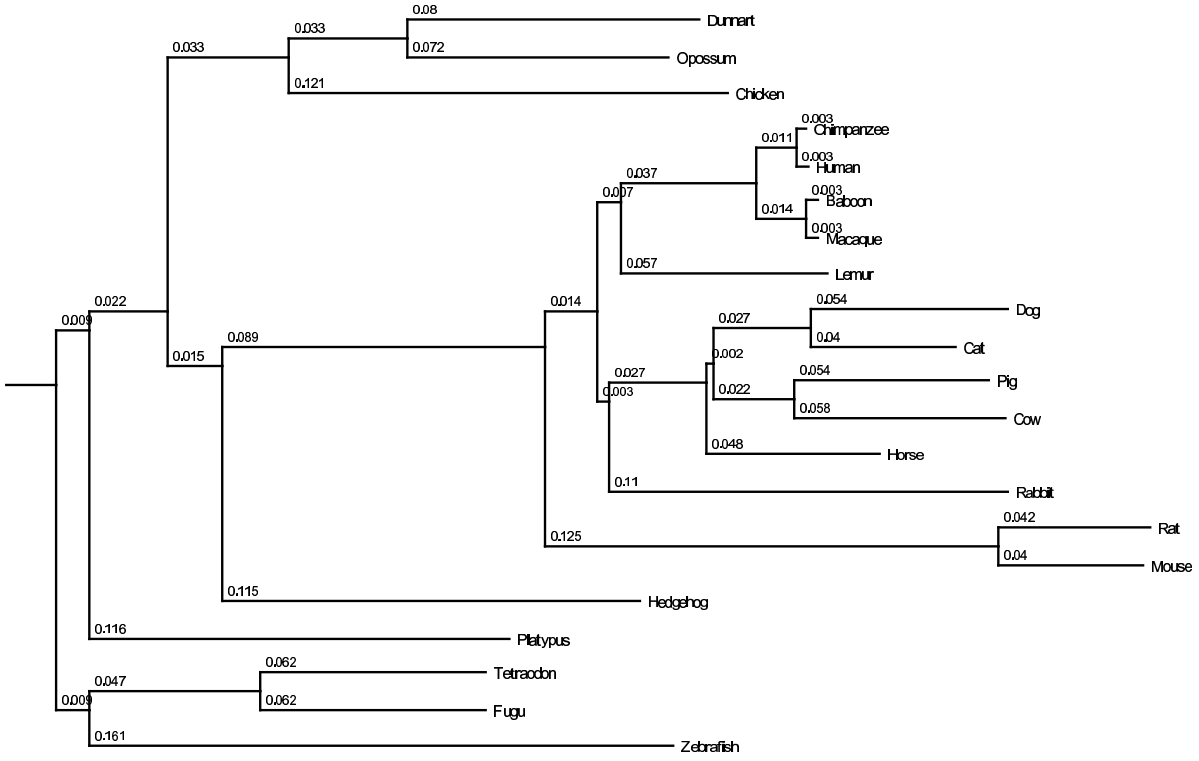

Figure 3: The 21-species phylogenetic tree estimate used in the empirical power analysis.